

Sparse Representation of Photometric Redshift PDFs: Preparing for Petascale Astronomy

Matias Carrasco Kind^{*} and Robert J. Brunner

Department of Astronomy, University of Illinois, Urbana, IL 61820 USA

28 April 2014

ABSTRACT

One of the consequences of entering the era of precision cosmology is the widespread adoption of photometric redshift probability density functions (PDFs). Both current and future photometric surveys are expected to obtain images of billions of distinct galaxies. As a result, storing and analyzing all of these PDFs will be non-trivial and even more severe if a survey plans to compute and store multiple different PDFs. In this paper we propose the use of a sparse basis representation to fully represent individual photo- z PDFs. By using an Orthogonal Matching Pursuit algorithm and a combination of Gaussian and Voigt basis functions, we demonstrate how our approach is superior to a multi-Gaussian fitting, as we require approximately half of the parameters for the same fitting accuracy with the additional advantage that an entire PDF can be stored by using a 4-byte integer per basis function, and we can achieve better accuracy by increasing the number of bases. By using data from the CFHTLenS, we demonstrate that only ten to twenty points per galaxy are sufficient to reconstruct both the individual PDFs and the ensemble redshift distribution, $N(z)$, to an accuracy of 99.9% when compared to the one built using the original PDFs computed with a resolution of $\delta z = 0.01$, reducing the required storage of two hundred original values by a factor of ten to twenty. Finally, we demonstrate how this basis representation can be directly extended to a cosmological analysis, thereby increasing computational performance without losing resolution nor accuracy.

Key words: galaxies: distance and redshift statistics – statistics – methods: data analysis – statistical

1 INTRODUCTION

Over the last two decades, photometric redshift (hereafter photo- z) estimation has become a crucial requirement for modern, multi-band photometric galaxy surveys. The need for accurate and robust photo- z s, and in particular for their probability density functions (photo- z PDFs), is increasing at an even faster rate as we enter the era of precision cosmology. Large photometric surveys like the Dark Energy Survey (DES¹) and the Large Synoptic Survey Telescope (LSST²) will probe hundreds of millions to billions of galaxies, which are often too faint to be spectroscopically observed. Even today, the Sloan Digital Sky Survey (SDSS; York et al. 2000) has obtained hundreds of millions of photometric observations of extragalactic sources covering approximately one-third of the sky. This same survey has also, by using a considerably larger amount of time with the same telescope, obtained a galaxy spectroscopic sample that is about one hundred times smaller. While the spectroscopic sample is a higher precision sample (Ahn et al. 2013), spectroscopy is both considerably

more difficult and more time consuming than photometry. For a number of cosmological analyses this higher precision is not strictly required, thus photometric surveys have become a cost-effective method for enabling new cosmological constraints.

As these photometric surveys have grown in popularity, the estimation of photo- z s that use different techniques has also grown significantly. These techniques can be broadly separated into machine learning algorithms (e.g., Collister & Lahav 2004; Carrasco Kind & Brunner 2013) or Spectral Energy Distribution (SED) fitting methods (see e.g., Benítez 2000; Ilbert et al. 2006). These two classes of techniques have been shown to have similar performance characteristics when high quality training data are available (see, e.g., Hildebrandt et al. 2010; Abdalla et al. 2011; Sánchez et al. 2014, for a review on some current photo- z techniques). More recently, particular attention has been focused on techniques that compute a full photo- z PDF for each galaxy. This is because a photo- z PDF contains more information than a single photo- z estimate, and the use of photo- z PDFs has been shown to improve the accuracy of cosmological measurements (e.g., Mandelbaum et al. 2008; Myers et al. 2009; Sheldon et al. 2012; Carnero et al. 2012; Jee et al. 2013) while not introducing any biases (e.g., Bordoloi et al. 2010; Abrahamse et al. 2011).

^{*} E-mail: mcarras2@illinois.edu

¹ <http://www.darkenergysurvey.org/>

² <http://www.lsst.org/lsst/>

One fact that all photometric surveys have in common is the need to efficiently handle an overwhelming quantity of imaging data. The reduction, analysis and storage of this data is a difficult problem; even with the growth of computational resources, efficiently handling these data remains a pressing problem. In particular, photo- z PDFs are currently computed on the summary catalogs that are produced by uniformly processing imaging data. But storing photo- z PDFs for billions of sources is a challenge in itself, which is further complicated if multiple, different photo- z techniques are desired or if different photo- z PDFs are generated by using different galaxy templates. This is a problem both for those managing the data archives and for the general community who desire to apply these photo- z PDF estimates to cosmological analyses. Thus the time is ripe to address this issue. In this paper, therefore, we explore different methods that allow us to manipulate and use photo- z PDFs in a more efficient manner by representing them as compactly as possible.

In this paper we introduce the use of a sparse functional basis to represent a full photo- z PDF. This approach minimizes the data required to represent the photo- z PDF, while maximizing the accuracy of the PDF. This basis representation not only minimizes the storage requirements, but also allow us to manipulate PDFs in a more computationally efficient manner, thereby increasing the computational efficiency of resulting analyses. With this approach, each galaxy photo- z PDF is decomposed into an over determined basis system by minimizing the number of basis functions retained. We analyze how this approach compares with other representation techniques, in particular with a multi-Gaussian approach; and we demonstrate that, by using our proposed functional form, the integration and manipulation of photo- z PDFs is both easier and faster than when using either the original PDF or any other comparable technique.

The rest of this paper is organized in the following manner. In Section 2 we first present the data we will use to test our proposed approach, after which we will introduce the photo- z methods that we will use to generate the photo- z PDFs that will be used in the rest of the paper. Section 3 describes the different algorithms used to represent the photo- z PDFs. In Section 4 we present the main results, compare the performance of the different representation methods, and discuss the storage requirements of these photo- z PDFs. Next, we apply our basis representation to the computation of $N(z)$ in Section 5, and discuss how this basis framework can reduce computational costs. We conclude in Section 6 with a summary of our main findings and a final discussion of the applications of our proposed approach.

2 DATA AND PHOTO-Z ESTIMATION

In this section, we first present a brief summary of the data that will be used to test different photo- z PDF representation techniques, including our proposed sparse representation basis. Next, we will present a short overview of TPZ, the photo- z PDF technique we use to generate the photo- z PDFs that are analyzed in this paper.

2.1 CFHTLenS

We use data from the Canada-France-Hawaii Telescope Lensing Survey (Heymans et al. 2012; Erben et al. 2013), here-

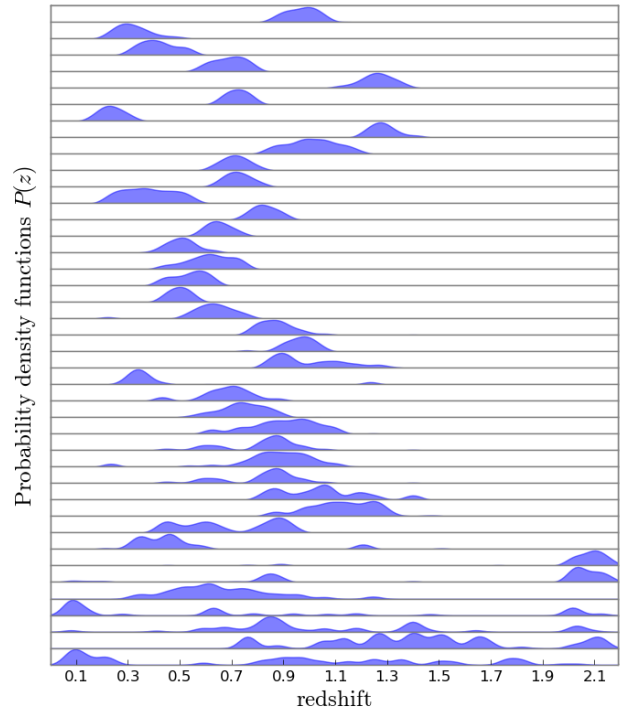


Figure 1. Forty representative, randomly selected photo- z PDFs computed for the CFHTLenS data by using TPZ, each normalized to unity. In each subplot, the horizontal axis is redshift and the vertical axis is the probability density. The PDFs are sorted in the vertical axis by the number of peaks.

after referred to as CFHTLenS³, with the photometry presented in Hildebrandt et al. (2012). This galaxy survey includes more than twenty-five million objects observed in five photometric bands: u , g , r , i , and z , covering 154 square degrees that includes all five years worth of data from the Wide, Deep and Pre-survey components of the CFHT Legacy Survey (CFHTLS; Gwyn 2012). To generate a spectroscopic training sample, we have cross matched the galaxies from the CFHTLenS with spectroscopic surveys whose sky coverage overlaps the four fields of the CFHTLenS survey.

In particular, we have selected high quality spectroscopic galaxies from the Deep Extragalactic Evolutionary Probe Phase 2 (DEEP2; Davis et al. 2007; Newman et al. 2013a), the VIMOS (VIsible imaging Multi-Object Spectrograph) VLT (Very Large Telescope) Deep Survey (VVDS; Le Fèvre et al. 2005; Garilli et al. 2008), the VIMOS Public Extragalactic Redshift Survey (VIPERS; Garilli et al. 2014), and the Sloan Digital Sky Survey Data Release 10 (Ahn et al. 2013, SDSS-DR10), which includes over two million spectra of galaxies and quasars taken as a part of the the Baryonic Oscillation Spectroscopic Survey (BOSS) program (Dawson et al. 2013). In the end, we have 49,868 high quality spectroscopic galaxies with a mean redshift of 0.6 to train our TPZ algorithm. As the objective of this paper is not to present photometric redshift PDFs for all of the CFHTLenS galaxies, we have randomly selected a subsample of 10^6 galaxies with no spectroscopic information from the survey that we use for the tests described within this text.

³ <http://www.cfhtlens.org/>

2.2 TPZ: Trees for Photo- Z

TPZ (Carrasco Kind & Brunner 2013) is a parallel, supervised machine learning algorithm that uses prediction trees and random forest techniques to produce both robust photo- z PDFs and ancillary information for a target galaxy sample. A prediction tree is built by *asking* a sequence of questions of the data. This process recursively splits the input data taken from the spectroscopic sample, frequently into two branches, until a terminal leaf is created that meets a stopping criterion (e.g., a minimum leaf size or a variance threshold). The dimension in which the data is divided is chosen to be the one with highest information gain among the random subsample of dimensions obtained at every point. This process produces less correlated trees and allows us to explore several configurations within the data. The small region bounding the data in the terminal leaf node represents a specific subsample of the entire data with similar properties. Within this leaf, a model is applied that provides a fairly comprehensible prediction, especially in situations where many variables may exist that interact in a nonlinear manner, as is often the case with photo- z estimation.

TPZ is publicly available through a software package called MLZ⁴ (Machine Learning for photo- Z). The full software package includes, among other capabilities, TPZ, a supervised learning method, and SOM z , an unsupervised photo- z technique (Carrasco Kind & Brunner 2014). As we have demonstrated in Carrasco Kind & Brunner (2014b), these techniques can be combined to make even more accurate photo- z PDF predictions. For simplicity, in this paper we restrict our analysis to TPZ only, since we focus within this paper solely on the accurate reconstruction of galaxy photo- z PDFs.

We compute the photo- z PDF for all one million galaxies in our test sample, described previously, by using our spectroscopic training sample. We used all colors and magnitudes, which results in a total of nine attributes, and construct 600 trees to make the predictions. TPZ also uses the attribute errors during the prediction process, in part to deal with missing attributes in the catalog (see Carrasco Kind & Brunner 2013, for a description of this approach). We also have computed photometric redshifts for galaxies in the training sample by using a cross validation technique called Out-Of-Bag (Breiman 2001; Carrasco Kind & Brunner 2013) in which a photo- z PDF is obtained for all galaxies in the training set by using all the trees that do not contain that particular galaxy. This approach, therefore, avoids over-fitting; and we have shown that this method is reliable and also unbiased (Carrasco Kind & Brunner 2014).

For illustration, we present a sample of forty photo- z PDFs randomly selected from the CFHTLenS galaxies in Figure 1, presented in increasing order by the computed mean value of their photo- z PDF. The redshift range for the galaxies are the same and the PDFs have all been normalized to unity. From this figure, it is clear that these photo- z PDFs are not simple functions, often having multiple peaks; and they are, therefore, poorly represented by a single Gaussian, which has often been used for simplicity in the past. In Figure 2, we present a summary of the results on the training data determined by using the OOB cross-validation technique. The top panel compares z_{phot} , computed by using the mean value of each photo- z PDF, with z_{spec} for all 49,868 galaxies in the training sample. This indicates the approximate performance

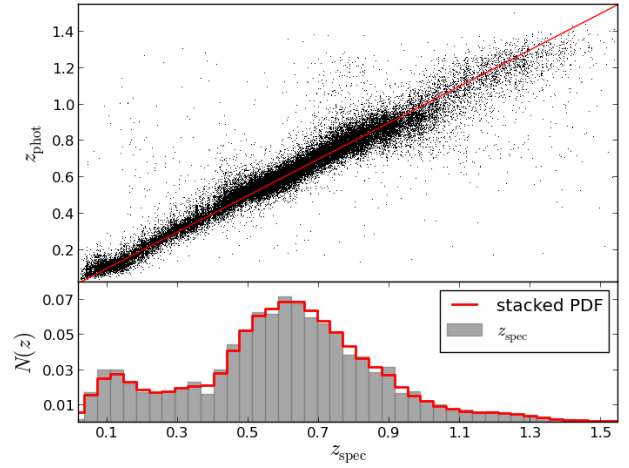


Figure 2. *Top:* A comparison of the photometric and spectroscopic redshift for all galaxies in the training sample computed by using the OOB cross-validation technique. The red line shows the one-to-one line for guidance. *Bottom:* The $N(z)$ distribution computed by using the spectroscopic redshifts (gray area) and by stacking the photo- z PDFs (red) for the training sample galaxies.

of TPZ on the real dataset within the limits of the training sample. The bottom panel shows the distribution of the galaxies as a function of redshift in terms of their spectroscopic values (in gray) compared with the $N(z)$ obtained by stacking the PDFs (red line). We can see a remarkable agreement between these two distributions, as we have shown previously (Carrasco Kind & Brunner 2013).

3 PDF REPRESENTATION

In this section we present the different methods that we use to represent the full photo- z PDF. For the rest of this discussion, we will make the following assumptions. First, we have N total galaxies in our sample with individual photo- z PDF estimates. Second, we can represent the photo- z PDF, $P_k(z)$, for the k^{th} galaxy in the sample by $\mathbf{p}z_k$. Finally, we have n sample points in the original galaxy photo- z PDF. Thus $P_k(z)$ is sampled at a resolution δz , given by $\delta z = \Delta z/n$, where Δz is the redshift range spanned by the photometric data.

3.1 Statistical Representation

The simplest representation for a full photo- z PDF is to use a summary statistic. We will consider five different, summary statistics, four of which are single values: the mean, the mode, the median of the PDF, and a Monte-Carlo sampling from the original cumulative PDF (Wittman 2009). This fourth approach involves sampling a random number from the range 0–1, and determining the cumulative probability to this numerical value. Finally, the last representation is a single Gaussian fit to the original photo- z PDF, which provides a two value summary: the mean and variance of the Gaussian. As a result, we require either N , for the first four approaches, or $2N$, for the last approach, statistics to represent the full photo- z PDF catalog.

⁴ <http://lcdm.astro.illinois.edu/code/mlz.html>

3.2 Multi-Gaussian Fitting

The second representation for a full photo- z PDF we explore is the application of a multi-Gaussian fit to each photo- z PDF (see, e.g., Bovy et al. 2011, 2012). In this approach, each Gaussian (when more than one is used) included during the fitting process will require three parameters: the amplitude, the mean, and the variance. In this approach, we first determine the number of peaks, Np_k , in the photo- z PDF. We increase this value by one and use the result as the number of Gaussians to be used in the fitting process for that specific photo- z PDF. The extra Gaussian improves the fit to extended wings in the photo- z PDF distribution, which often arise from the residuals of the Gaussian fits to the individual PDF peaks.

To determine the best fit values, we use a Levenberg-Marquardt minimization algorithm. In this case, each $P_k(z)$ can be represented by:

$$\mathbf{pz}_k = \sum_{i=1}^{Np_k+1} \alpha_{k,i} e^{-\frac{(\mathbf{z}-\mu_{k,i})^2}{2\sigma_{k,i}^2}} \quad (1)$$

where α_k , μ_k and σ_k^2 are vectors of dimension $Np_k + 1$ containing the amplitude, mean, and variance for each Gaussian included in the fitting process. As a result, the total number of values needed to represent the full photo- z PDF catalog is $\sum_k 3(Np_k + 1)$.

3.3 Sparse Basis Representation

The final technique that we will use to represent a photo- z PDF is the sparse basis representation. In this case, we will adopt a set of basis functions to represent a PDF by using the following model:

$$\mathbf{pz}_k = \mathbf{D}\boldsymbol{\delta}_k + \boldsymbol{\epsilon}_k \quad (2)$$

where \mathbf{D} is a dictionary or basis matrix of dimension $n \times m$, where $m > n$. Thus, we have an over-determined problem as the number of basis functions, m , is much larger than the dimension, n , of each photo- z PDF. In this case, each column, \mathbf{d}_j , of the dictionary matrix, \mathbf{D} , is a basis function that must be ℓ_2 normalized, i.e., $\|\mathbf{d}_j\|_2 = 1$ for $j = 1, 2, \dots, m$.

We want to find, for each galaxy k , the optimal vector solution $\boldsymbol{\delta}_k$, which is determined such that its pseudo-norm $\|\boldsymbol{\delta}_k\|_0$ is minimized. Alternatively, this can be equivalently stated that we want to minimize the number of non-zero entries in the vector, $\boldsymbol{\delta}_k$, given the residual error $\boldsymbol{\epsilon}_k$. In this case, we can either use a predefined number of basis functions or we can define a fixed residual for every galaxy in the sample. Either way, the total number of points required to represent the entire catalog is given by $\sum_k 2(Nb_k)$, where Nb_k is the number of basis functions used for each galaxy. Note that in this case we only need two numbers for each functional basis: the functional coefficient (a floating point number), and the index number of the function within the basis set (a integer number). This already corresponds to a potentially large reduction in the total data volume required to archive photo- z PDFs. We will see in subsequent sections that we can also represent the basis function coefficients by using integers; and that, in addition, we can combine both terms into a single thirty-two bit integer, thereby reducing the total number of values required to $\sum_k (Nb_k)$.

Finding $\boldsymbol{\delta}_k$ in this over determined problem can be challenging. For this analysis, we have selected to use Orthogonal Matching Pursuit (OMP), an iterative algorithm that finds,

at each step, the column, \mathbf{d}_j , of the dictionary matrix, \mathbf{D} , that best represents the current residuals. This process is repeated until a predefined criteria is reached, either a residual threshold or the total number of basis functions used. Fundamentally, this approach is similar to the well known CLEAN algorithm, which is used to analyze interferometric radio observations (Högbom 1974). The advantage of OMP over the standard Matching Pursuit algorithm (Mallat & Zhang 1993) is that a specific basis function can only be selected once. Since the residuals are orthogonalized during the selection of the basis functions for the current galaxy, we generate an independent set of basis functions to represent each galaxy's photo- z PDF.

Conceptually, the OMP algorithm that we apply to all galaxies can be enumerated⁵:

1) Initialize all variables. First, define the residual vector to be the original photo- z PDF, $\boldsymbol{\epsilon}_k^0 = \mathbf{pz}_k$. Second, create an empty set of cumulative selected basis functions, \mathbf{B}_k . Finally, set $\boldsymbol{\delta}_k = 0$, and define $i = 0$ as the number of the current iteration.

2) Compute the current set of basis functions. First, find the column vector, \mathbf{d}_b , from the dictionary matrix, \mathbf{D} , where b is the index position that maximizes the projection of $\boldsymbol{\epsilon}_k^i$:

$$\mathbf{d}_b^i = \max_{\mathbf{d}_j \in \mathbf{D}} |\mathbf{d}_j^T \cdot \boldsymbol{\epsilon}_k^i| \quad (3)$$

Second, add this selected basis function to the set \mathbf{B}_k , i.e., $\mathbf{B}_k = (\mathbf{B}_k, \mathbf{d}_b^i)$.

3) Orthogonally project the original photo- z PDF onto the linear space spanned by the columns of all previously selected basis functions:

$$\mathbf{w}_k^i = \mathbf{B}_k^T \cdot \mathbf{pz}_k \quad (4)$$

where \mathbf{w}_k^i is a temporary vector corresponding to the coefficients of the currently used basis functions in \mathbf{B}_k .

4) Complete the projection by updating the residuals by using the temporary vector \mathbf{w}_k^i :

$$\boldsymbol{\epsilon}_k^{i+1} = \mathbf{pz}_k - \mathbf{B}_k \cdot \mathbf{w}_k^i \quad (5)$$

5) Check the stopping criteria: $\|\boldsymbol{\epsilon}_k^{i+1}\|_2 < \epsilon_{th}$, where ϵ_{th} is the threshold residual or $i > i_{lim}$, where i_{lim} is the number of required basis functions. If the pre-selected stopping criteria is met, the calculations are completed: $\boldsymbol{\delta}_k = \mathbf{w}_k^i$ and $\mathbf{pz}_k = \mathbf{D} \cdot \boldsymbol{\delta}_k + \boldsymbol{\epsilon}_k^{i+1}$, where $\boldsymbol{\delta}_k$ is sparse. Finally, the photo- z PDF representation is defined:

$$\mathbf{pz}_k \approx \mathbf{D} \cdot \boldsymbol{\delta}_k \quad (6)$$

On the other hand, if the predefined stopping criteria is not met, the iteration step is increased, $i = i + 1$, and steps 2–5 are repeated by using the current residual vector. This process is repeated over all galaxies k , where $k = 1, 2, \dots, N$.

Dictionary Selection

Given the nature of the shape of the photo- z PDFs (see, e.g., Figure 1), it is natural to select a set of Gaussian-like basis functions that span the redshift range of our photometric galaxy sample. We can use the the original resolution and redshift range spanned by the generated photo- z PDF to determine the dictionary to use for the sparse basis representation.

⁵ Note, the superscript T indicates transposition

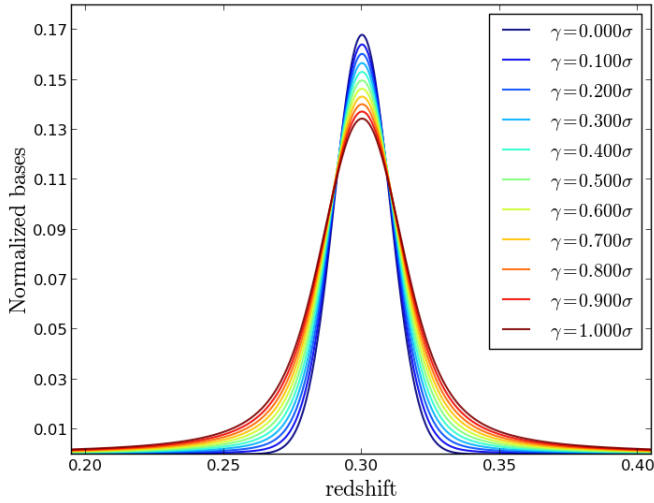


Figure 3. Different normalized $\|\mathbf{d}_j\|_2 = 1$ Voigt profile basis functions with the same mean, $\mu = 0.3$, and sigma, $\sigma = 0.01$, for different values of γ , which ranges from 0 (blue) to 1σ (red). Note that for $\gamma = 0$, we recover the standard Gaussian distribution. In a full dictionary, we create these profiles over the entire redshift range of the galaxy sample for different values of σ .

One of the primary advantages of this method is that these dictionary entries are composed of analytic functions that can be combined with other functional forms. There are no restrictions, other than computational time, on how large of a dictionary we can use, as there is no requirement for the dictionary to be permanently stored. Furthermore, a photo- z PDF can be restored even without reconstructing the dictionary, as long as the indices and coefficients are efficiently stored.

We select N_μ Gaussian functions, whose mean values span the redshift range of our galaxy sample, which has a redshift resolution δz . Thus, we can compute:

$$N_\mu = \left\lceil \frac{\Delta z}{\delta z} \right\rceil \quad (7)$$

where $\Delta z = z_2 - z_1$ and z_2 and z_1 are, respectively, the upper and lower limits of the redshift range spanned by our galaxy sample. We select, at each N_μ location, N_σ values for the standard deviation that linearly span the range from a minimum value of σ_{\min} to a maximum value σ_{\max} . The minimum value is selected in such a way that we will approximately have a single Gaussian that fills a single redshift bin of width δz . In practice, a Gaussian vanishes at approximately 3σ from the mean; therefore, we can select $\sigma_1 = \delta z/6$.

On the other hand, we select the broadest basis function to approximately cover half of the full redshift range Δz at each position; therefore, we select $\sigma_{\max} = \Delta z/12$. Although the extreme basis functions are not frequently used, they ensure that we cover all possibilities. Finally, we set the resolution between different values of σ to be $\delta z/2$ in order to make sure the difference between two consecutive Gaussian basis functions is on the order of δz . Setting $\Delta\sigma = \sigma_{\max} - \sigma_{\min}$ we have that N_σ is given by:

$$N_\sigma = \left\lceil \frac{2\Delta\sigma}{\delta z} \right\rceil \quad (8)$$

which can be simplified to

$$N_\sigma = \left\lceil \frac{\Delta z}{6\delta z} - \frac{1}{3} \right\rceil \approx \frac{N_\mu}{6} \quad (9)$$

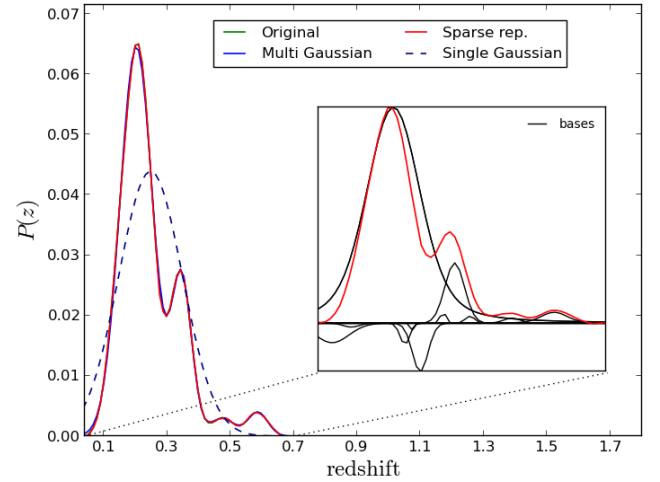


Figure 4. The representation of an original photo- z PDF (green) given by three techniques: multi-Gaussian (blue), single Gaussian (blue dashed line), and sparse basis representation (red). The inset panel shows the final bases (in black) used to represent the photo- z PDF while the recovered distribution is shown in red.

As some photo- z PDFs have extended wings, we also generate N_γ basis functions for each Gaussian basis function with extended profiles by using a Voigt profile. Voigt profiles are widely used in spectral line fitting, and are defined as the convolution between a Gaussian distribution and a Lorentzian distribution. A Voigt profile can be written as the real part of the Faddeeva function (Abramowitz & Stegun 1972):

$$V(x; \sigma, \gamma) = \frac{1}{\sigma\sqrt{2\pi}} \operatorname{Re} \left[e^{-z^2} (1 - \operatorname{erf}(-iz)) \right] \quad (10)$$

where $\operatorname{erf}(-iz)$ is the complex *error function*. $z = \frac{(x-\mu)+i\gamma}{\sigma\sqrt{2}}$ is a complex variable, where μ is the center of the function, σ is the standard deviation from the Gaussian, and γ determines the strength of the extended wings and is a parameter from the Lorentz distribution. As a result, if $\gamma = 0$, we have a Gaussian distribution with parameters μ and σ .

We present examples of different Voigt profiles in Figure 3 given a fixed $\mu = 0.3$ and $\sigma = 0.01$, but with γ varying from zero (Gaussian) to one σ . We do not, however, select pure Lorentzian profiles, as they produce distributions that are too extended to be practical for this analysis. In practice, we find that an upper limit of $\gamma = 0.5\sigma$ is sufficient to accurately model any extended wings. Thus, including the Gaussian case with $\gamma = 0$, we fix $N_\gamma = 6$ and allow γ to vary linearly from 0 to 0.5σ in steps of 0.1σ . Thus, in the most simple case we would only consider basis functions with $\gamma = 0$ and $N_\gamma = 1$.

In total, the dictionary is composed of $N_{\text{total}} = N_\mu \times N_\sigma \times N_\gamma$ bases, which all have ℓ_2 norm equal to unity. By using our previous definitions, we have the following approximate rule of thumb for creating a dictionary:

$$N_{\text{total}} \approx N_\mu^2 = \left(\frac{\Delta z}{\delta z} \right)^2 \quad (11)$$

Although this is an estimate, it provides a very good approximation to the total number of bases needed given the resolution of the original photo- z PDF. Additional bases are not necessary and little is gained by using a finer resolution. Photo- z codes generally provide photo- z PDFs by using roughly two to three hundred points. According to Equation 11, we notice that for 250 sample points in a PDF, we would need approx-

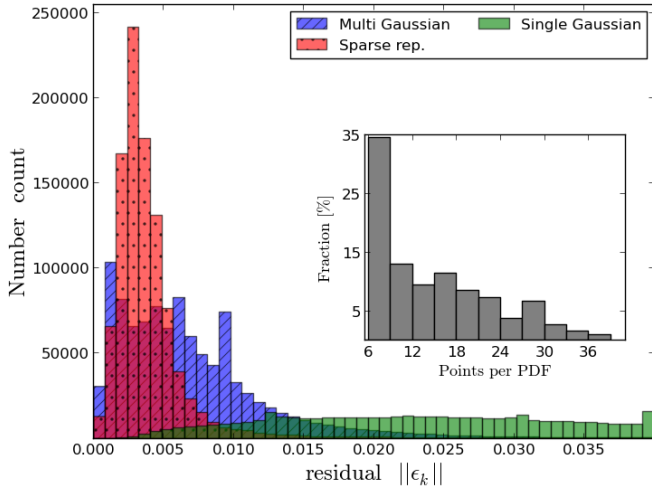


Figure 5. The residual distribution for all CFHTLenS galaxies computed by using the multi-Gaussian representation (blue) and the sparse basis representation (red). In each case, we use the same number of representation values for each galaxy photo- z PDF. For comparison, the Single-Gaussian representation is shown in green. *Inset:* The distribution of points (bases or fitting parameters) per galaxy photo- z PDF. The number of peaks, Np_k , per photo- z PDF is the same, but divided by 6. Thus, there are $3(Np_k + 1)$ parameters per galaxy.

imately 62,500 bases. Thus, we can use a 2-byte integer to express the indices into our basis function dictionary, which has important ramifications in the compact storage of photo- z PDFs as discussed in the next section.

4 DISCUSSION

We have applied the previously discussed photo- z PDF representation techniques to the CFHTLenS data introduced in Section 2. We have computed a photo- z PDF for each galaxy in the one million test sample by using the TPZ software to compute a PDF with two hundred sampled points at a resolution of $\delta z = 0.011$. We display one such photo- z PDF in Figure 4 where the original distribution is shown in green, a multi-Gaussian representation is shown in blue, a sparse basis representation is shown in red, and a single Gaussian model is shown with a blue dashed line. We can see that both the sparse basis representation and the multi-Gaussian agree remarkably well with the original photo- z PDF, to the point where it is hard to see the original PDF. As one would expect in this multi-peak PDF, the single Gaussian model does not reproduce this photo- z PDF very well. The inset panel provides a zoomed-in view showing the sparse basis representation of the photo- z PDF and the actual basis functions used in the representation. As the number of bases is increased, we expect some of them to have a negative coefficient, as shown in the inset, which aids in the reconstruction of the residuals from the previous bases. Given the iterative nature of this process, we select the new basis function that optimally corrects the residuals of previous bases in order to best reconstruct the photo- z PDF by using the minimum number of functions.

In order to quantitatively compare the reconstruction of the photo- z PDF by using the three methods as shown in Figure 4, we compute the multi-Gaussian fitting for all 10^6 galaxies from our CFHTLenS test sample. For each galaxy we

record the number of values (parameters) required to accurately reconstruct the original PDF. Note that in this fitting approach, we are not fixing the number of Gaussian functions used in the reconstruction, but are instead defining the number of Gaussians as the number of peaks in the photo- z PDF plus one extra Gaussian to compensate for the residuals and extended profiles. In addition, we also compute, for each galaxy, the optimal sparse representation by using a variable number of basis functions that are constrained to match the number of points used in the multi-Gaussian fitting. We also compute the best single Gaussian fit to each PDF to demonstrate the importance in using the information contained within the full PDF as opposed to simply treating each photo- z estimate as a Gaussian PDF.

After computing the different representations for each galaxy, we next compute the norm of the residuals between each representation and the original photo- z PDF for each galaxy and accumulate the results. We compare the resulting distributions in Figure 5. First, we notice the broad shape of the single Gaussian distribution (green). In fact, the width of the distribution exceeds the plot boundaries as the median of the single Gaussian distribution of residuals is 0.043, which is outside the range of the Figure. Second, we observe that when using the same number of values to represent the photo- z PDF, the sparse basis representation produces much smaller residuals with a more concentrated distribution than the multi-Gaussian fitting. Specifically, the median of the sparse representation residual distribution is 0.0033, which is almost half of the value (0.0058) for the multi-Gaussian fitting. Both of these results indicate that either method provides a good representation of the photo- z PDF by using a small number of values. We also show the distribution of values required to reconstruct the photo- z PDF of each galaxy in the inset panel of Figure 5. This subplot indicates that approximately 35% of the galaxies are single peaked (six values are required for two Gaussians, which in our implementation means a single peak plus an extra Gaussian for the extended wings). The distribution extends up to thirty-nine values for roughly 1-2% of the sample, which corresponds to twelve peaks in a photo- z PDF. The average number of values per galaxy is fourteen, which, in itself, implies a large compression ratio when compared to the original two hundred values while still providing a very good reconstruction of the full photo- z PDF.

While a natural number of basis functions can be determined for the multi-Gaussian representation, the sparse basis representation is more general and thus does not have a simple, natural number of basis functions. In order to better understand the optimal number of basis functions for photo- z PDFs, we compute the sparse basis representation for all galaxies in the test sample by using a different number of fixed bases. We combine the residuals, and plot the median value of the distribution as a function of the number of values used to represent the PDF as blue dots in Figure 6. As shown in the figure, as we increase the number of bases, the residuals decrease monotonically. This decrease is quite rapid at first, as expected, and slowly decreases until approximately twenty-five bases are used. For comparison, we also show the multi-Gaussian residuals for fourteen values (red triangle) and the corresponding sparse basis representation residuals for approximately the same number of values (black star), demonstrating the superiority, in terms of precision, of the sparse representation over the multi-Gaussian. If we restrict the number of values to twenty, we have a median residual of 0.018,

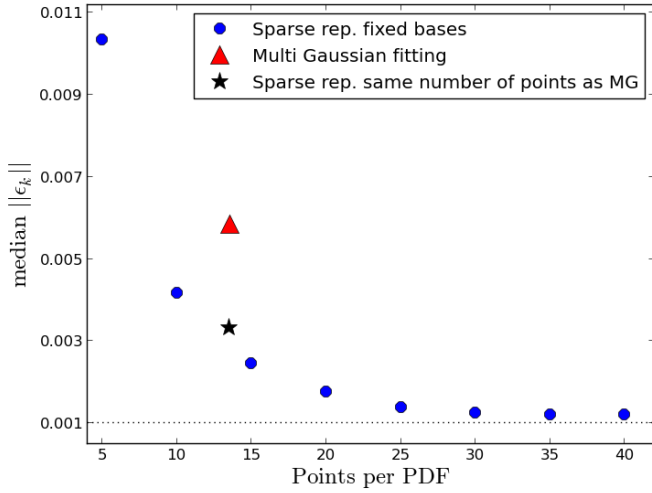


Figure 6. The median of the residual distribution as a function of the number of fixed bases used to reconstruct each galaxy’s photo- z PDF when using the sparse representation technique (blue dots). For reference, the median of the multi-Gaussian residual distribution (red triangle) and the median of the sparse representation with variable number of bases (black star) are also shown, where on average both techniques need fourteen points per photo- z PDF.

which corresponds to a median reconstruction of all one million test galaxies at 99.82% at a resolution of $\delta z = 0.011$. Since the original photo- z PDF contained two hundred points, this implies a compression ratio of ten.

Clearly these results will vary depending on the galaxy sample. In particular, the data we use in this analysis are from the CFHTLenS, which is a representative deep survey with galaxies that have photo- z PDFs with up to twelve peaks. The performance of the sparse representation also depends directly on the number of peaks in each PDF when we globally fix the number of bases. In Figure 7, we display the median of the residual distribution as a function of the number of peaks in the photo- z PDF, with different curves corresponding to different numbers of globally fixed bases. For a fixed number of bases, the residual increases as the number of peaks increase. Thus, a galaxy sample that consistently has a low number of peaks will have increased performance when using a smaller number of bases.

For example, we achieve a 99.5% reconstruction by using only ten values for galaxies with four or fewer peaks. In Carrasco Kind & Brunner (2014b), we discussed the relationship between the number of peaks and the shape of the photo- z PDFs with the outlier fraction. With this in mind, we could reduce the number of bases used to reconstruct a sample and flag those with a high number of peaks, where the reconstruction is less reliable, for further investigation. In fact, we achieve a reconstruction of 99% for photo- z PDFs with three or fewer peaks when using only five bases for the sparse representation. This produces a compression ratio of forty when the original photo- z PDF has two hundred points.

For comparison, we also show the fitting residuals for the multi-Gaussian (black dashed line) and sparse representation (black dashed-dotted lines) where the variable number of bases matches the number of multi-Gaussians. The performance of the multi-Gaussian fitting is less dependent on the number of peaks simply because the number of parameters dynamically changes for each photo- z PDF. Overall, the multi-Gaussian performance is fairly consistent at around 0.005, even as we

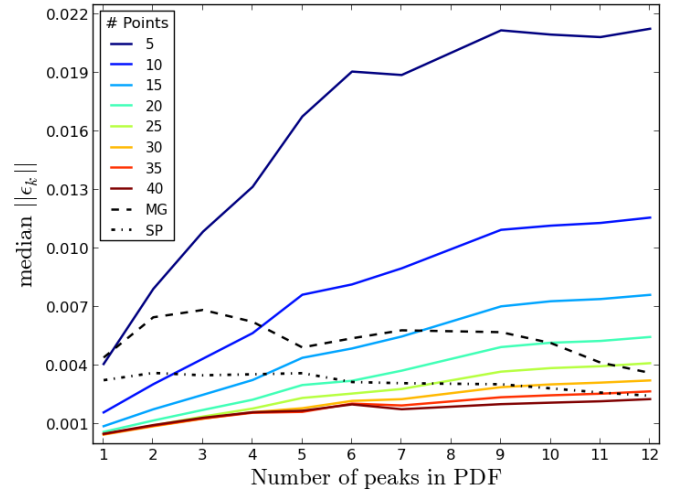


Figure 7. The median of the residual distribution as a function of the number of peaks in the photo- z PDF when using (solid color lines) a different number of fixed bases in the sparse basis representation, (black dashed line) when using the multi-Gaussian fitting technique, and (black dashed-dotted line) when using the sparse representation when the number of bases is equivalent to the number of multi-Gaussians.

increase the number of peaks. The sparse representation with a variable number of bases, on the other hand, is less dependent on the number of peaks and has residuals that are nearly 50% smaller than the multi-Gaussian fitting at an approximately constant value of 0.003.

PDF Storage

In the previous section, we discussed how the sparse representation and the multi-Gaussian fitting can accurately represent a photo- z PDF by using only a few dozen values with a reconstruction level of 99%. In the case of the multi-Gaussian fitting, the number of parameters to be stored will depend on the number of peaks in each individual PDF. As discussed previously, we will have $3(Np_k + 1)$ parameters, which are all floating point numbers. For this dataset we found that the average number of values (or floating point parameters) required is fourteen; but to store these data for all galaxies, we would need to combine the results from different galaxies in order to take advantage of the galaxies that require fewer values so that we can also store those galaxies that require a larger number of parameters. Varying the number of values to store galaxy photo- z PDFs in this manner might not be practical, as it will likely depend strongly on the archival and storage system while also increasing the computational difficulty in dealing with a varying number of parameters for different photo- z PDFs. The practical solution would be to use thirty-nine fixed values (the maximum required for this dataset) for all galaxies and store them independently. This result is also true for the varying sparse representation, which we have demonstrated has a better performance in comparison to the multi-Gaussian when representing a photo- z PDF.

On the other hand, requiring a fixed number of basis functions per galaxy alleviates this issue and also has the additional benefit that there is no need to pad with zeros since having more points for single peaked galaxies simply provides a more accurate representation. We have shown that by using ten to twenty values we are able to produce a residual on the or-

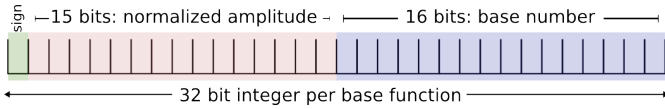


Figure 8. A single four-byte integer scheme to store a single basis function in the sparse representation method. The first sixteen bits store the coefficients (including sign), while the second sixteen bits store the location of the bases in the dictionary.

der of 0.1%, where all galaxies are stored independently. One additional (and very important) advantage of the sparse basis representation is that all bases in the dictionary have ℓ_2 norm equal to unity. Furthermore, when bases are computed by using the OMP algorithm, the absolute values of all coefficients are, by definition, less than unity. They can be negative, however, as seen in Figure 4. Since the PDFs are probability distributions, by definition the integral of the PDFs over the redshift range must also be unity. As a result, we can rescale all coefficients; and, as long as their relative amplitudes are the same, we can always impose the integral normalization at the end of the reconstruction.

If we continue this line of reasoning, we can rescale the coefficients of every basis function for a given galaxy so that the coefficients have absolute values between zero and one. When doing this we will be sure that the first basis function has unit amplitude without loss of accuracy on the very first basis. We can discretize this range by using approximately 32,000 sampling points (specifically 2^{15}) between zero and one, and store the corresponding integer from this range, and its sign, in a single sixteen-bit value. The error introduced by this discretization is very small, on the order of 10^{-5} , which is almost always negligible for most applications. In this approach, the most important basis is always first, and since it defines the scale, is always stored with no rounding errors.

We have, in fact, used this discretization throughout this paper; the difference introduced by using this discretization and the real values is less than 0.0005% and thus it does not directly affect the representation accuracy. This allows us to only use one value per basis function in our sparse representation. Since our dictionary contains fewer than 65,000 bases, which can be completely represented by a sixteen-bit integer, we can use a single four-byte integer, as shown graphically in Figure 8, to store both the base function and its amplitude and sign. More specifically, if we have a two hundred point photo- z PDF, which corresponds to a resolution of $\delta z = 0.011$ over the range $z = 0$ to $z = 2.2$, and we fix our representation to use ten bases, we can achieve an average reconstruction accuracy of 99.5% by using only 40Bs per photo- z PDF. Given a million galaxies that we treat in this manner, we will only need approximately 38MBs to store all of their PDFs. In addition, since we are only storing four-byte integers to represent the full photo- z PDF, we can potentially reduce the overall disk storage requirements by employing existing bit compression techniques (Lemire & Boytsov 2012) which will be important for relational database systems.

The representation and data encoding scheme we have proposed is, of course, even more flexible than we have demonstrated. If our photo- z PDFs employ either a different redshift resolution, span a larger redshift range, or simply have been sampled at a higher number of points, we can still use a four-byte integer representation. For example, if the original PDF is sampled at a finer resolution, we can double N_μ and reduce N_γ by one-half and still retain the same number of bases, re-

call we simply need the number of bases to be less than 2^{16} (or 65,536) in order to still have 2^{15} bits to encode the basis function index. In an extreme case, we can revert to a purely Gaussian set of basis functions and allow N_μ and N_σ to vary while keeping the total number of bases below the 2^{16} limit. In this case, we likely would need to increase the number of fixed bases in order to accurately represent the photo- z PDF.

If the number of required bases exceeds the 2^{16} limit, because, for instance, our photo- z PDFs are sampled at an extremely high resolution or span a large redshift range, we can always increase the size of the dictionary beyond this two-byte limit. In this case, we simply have a very dense dictionary, where fewer fixed bases would be necessary; thus, each basis function would be stored in either a six-byte or an eight-byte integer, depending on the details of the computational system. Another alternative would be to fix the number of bits used to encode each type of basis function; for example, to use two-bits for N_γ , six-bits for N_σ and eight-bits for N_μ , resulting in four, sixty-four, and two hundred and fifty-six possible values for each basis function. As a fixed framework, this technique could also simplify the storage and functional indexing. Finally, as we have mentioned earlier, there is no need to store the entire dictionary since it is simply defined over a functional basis. Instead, we only need to store the parameters required to regenerate the dictionary so that we can either regenerate the dictionary or generate the individual functions themselves as needed.

5 APPLICATION

We now change our focus from the computation of a compact representation of a photo- z PDF to a demonstration of their use in a scientific analysis. Perhaps the simplest application of a photo- z PDF is the computation of the number of galaxies as a function of redshift, which is widely used in the statistical analysis of the spatial distribution of galaxies. This function is computed by binning spectroscopic observations of galaxies as a function of redshift; but for a photometric survey, this distribution is optimally computed by integrating over all individual photo- z PDFs at a given resolution. This approach, while more computationally demanding, has been shown to be better than simply using a single estimate for the photo- z (e.g., Mandelbaum et al. 2008; Sheldon et al. 2012; Carrasco Kind & Brunner 2013). Even with this simple application, however, we benefit from the use of a sparse representation for our photo- z PDFs, since we can transform our theoretical framework to use our basis functions. Thus we can operate directly over the dictionary and use the sparse basis indices and coefficient parameters to calculate the true and reconstructed values taking into account the normalization.

As a demonstration, we can compute the galaxy redshift distribution directly over the basis functions. We start by writing the definition of $N(z)$:

$$N(z) = \sum_{k=1}^N \int_{z-\Delta z/2}^{z+\Delta z/2} P_k(z) dz \quad (12)$$

where the sum is over all N galaxies and $P_k(z)$ is the photo- z PDF of a given galaxy k . z is the midpoint of each redshift bin, which have a fixed width Δz . We can rewrite this equation in terms of the PDF representation for $P_k(z)$, which we previously defined as $\mathbf{p}z_k$. Thus, in the sparse basis representation,

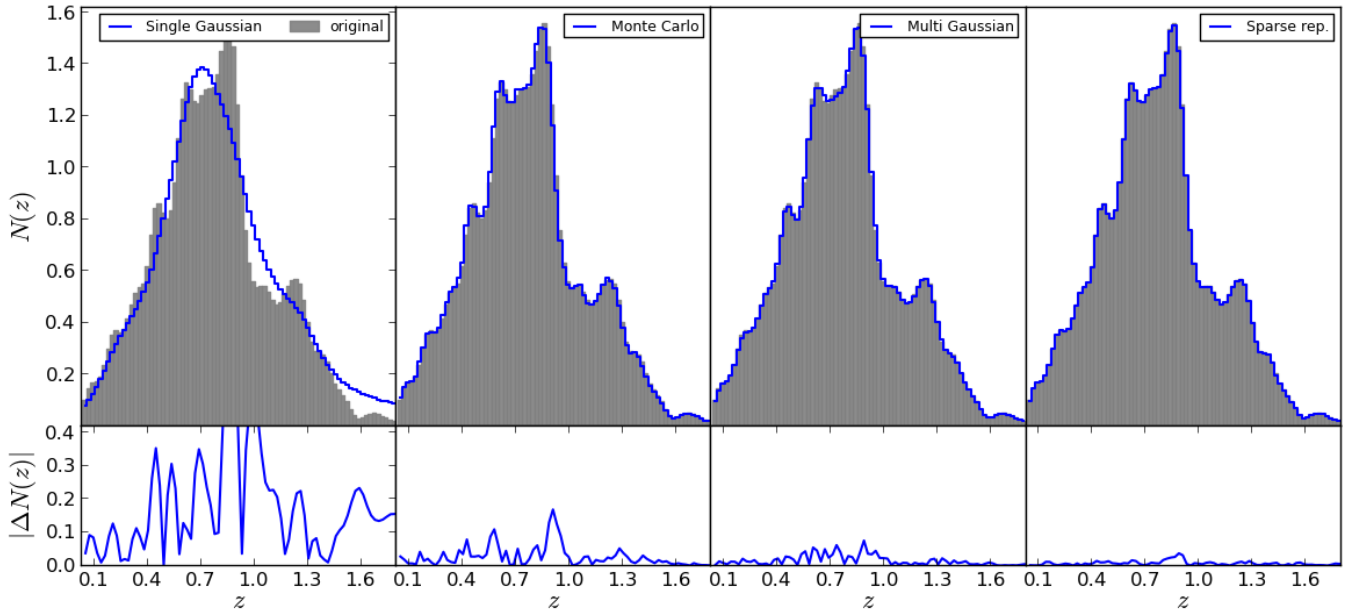


Figure 9. The $N(z)$ distribution for all 10^6 galaxies from the CFHTLenS data computed by using the four photo- z PDF representation techniques. Within each panel, the original $N(z)$ computed by stacking the full photo- z PDFs is shown in gray and a different representation method is shown in blue. From left to right we have the single Gaussian model, the Monte Carlo sampling described in Section 3.1, the multi-Gaussian fitting method, and the sparse representation method that uses the same number of bases as the multi-Gaussian method. The bottom panels show the absolute difference between the original and reproduction at the same scale.

we can express each PDF as:

$$\mathbf{p}z_k \approx \mathbf{D} \cdot \boldsymbol{\delta}_k \quad (13)$$

where $\boldsymbol{\delta}_k$ is a sparse vector, which might contain ten to twenty elements, that contains the amplitudes for each functional basis and \mathbf{D} is an $n \times m$ dictionary, where n is the number of points in the original PDF and m is the total number of bases. By using this result, we can rewrite Equation 12:

$$N(z) = \sum_{k=1}^N \boldsymbol{\delta}_k \cdot \int_{z-\Delta z/2}^{z+\Delta z/2} \mathbf{D} dz \quad (14)$$

where $\boldsymbol{\delta}_k$ is independent of redshift so that we only need to integrate once over each basis function in the dictionary; thus, we only have m integrations instead of N .

Furthermore, we can precompute this integral over \mathbf{D} , which we denote by $\mathbf{I}_D(z)$. This integral corresponds to a vector of length m , where each entry is the integral over each one of the basis functions \mathbf{d}_j in \mathbf{D} :

$$\mathbf{I}_D(z) = \int_{z-\Delta z/2}^{z+\Delta z/2} \mathbf{d}_j dz \quad j = 1, 2, \dots, m \quad (15)$$

Since all N galaxies are expressed in terms of the m bases, we can also pre-factorize the coefficients (or amplitudes) in a vector $\boldsymbol{\delta}_N$:

$$\boldsymbol{\delta}_N = \sum_{k=1}^N \boldsymbol{\delta}_k \quad (16)$$

Therefore, after these precomputations we can simply express $N(z)$ as:

$$N(z) = \mathbf{I}_D(z) \cdot \boldsymbol{\delta}_N \quad (17)$$

reducing the computation to a simple dot product of precomputed quantities. For each bin, we need to compute $\mathbf{I}_D(z)$, but $\boldsymbol{\delta}_N$ is computed only once and can be used both for all bins

in the computation of $N(z)$ and in other cosmological applications. This result is also true for other linear operations that might be involved in another cosmological analysis. Thus, by working directly in the space defined by the basis functions, we can reduce computational memory and processing times significantly.

We compare the original $N(z)$ for 10^6 test galaxies from the CFHTLenS sample to different $N(z)$ distributions reconstructed by using Equation 17 for different representation formats in Figure 9. Each original galaxy photo- z PDF has a resolution of $\delta z = 0.011$ and contains two hundred values. We restrict the comparison in Figure 9 to four techniques: a single Gaussian model, the Monte Carlo estimator described in Section 3.1, a multi-Gaussian fitting technique, and the sparse basis representation. However, we compute the fractional percentile error between the original $N(z)$ and all eight techniques and compare the results in Table 1. Before discussing the performance of individual techniques, we note that the lower panels in Figure 9 are all shown at the same scale to facilitate direct comparisons.

In the first panel, we see that the single Gaussian model clearly shows a significant difference, which is visible both from the distribution itself and in the bottom panel from the absolute error between these two distributions. Next, we see that the single point photo- z estimation computed by using a Monte Carlo sampling shows a surprisingly good agreement with the original distribution. This result was discussed by Wittman (2009), who demonstrated that this technique does provide a fair statistical representation of the sample's galaxy redshift distribution. This approach, where the $N(z)$ distribution is computed as a random sample drawn from the cumulative PDF of each galaxy, statistically compensates for the photo- z errors for an individual galaxy and thus produces a reliable $N(z)$ distribution. This approach does, however, introduce much larger errors on the estimation of individual galaxy

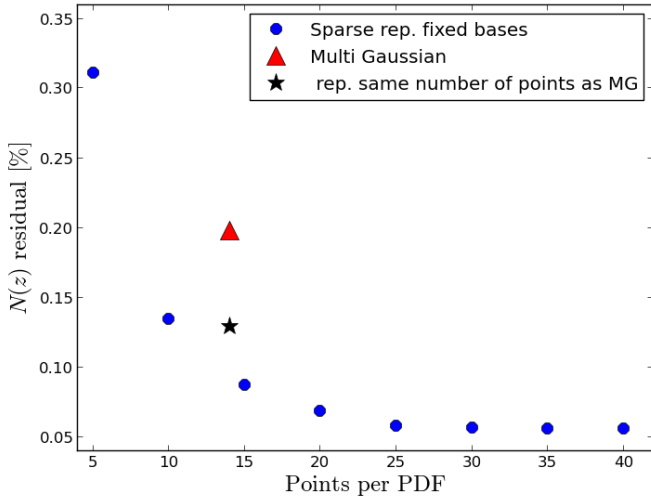


Figure 10. The fractional percentile error between the original $N(z)$ and the reconstructed $N(z)$ computed by using Equation 17 when fixing the number of bases used to represent the photo- z PDFs for all galaxies with the sparse representation technique (blue dots). For comparison, we also show the multi-Gaussian (red triangle) and the sparse representation with a variable number of bases (black star) residuals. On average, both of these latter techniques require fourteen points per photo- z PDF.

photo- z s. While one might be tempted to store a photo- z PDF by using this approach in order to accurately recover an $N(z)$ by storing a minimum quantity of new data, it would be easier to simply compute and store the actual $N(z)$. Furthermore, since this technique is dependent on using a large number of galaxies to generate a more accurate $N(z)$, if one is interested in computing the redshift distribution for galaxy subsets, the reconstruction accuracy might suffer as the number of galaxies in the subsamples is decreased.

In the third panel, we see that the multi-Gaussian fitting technique has a smaller error than the first two methods. As discussed previously, this method provides an accurate representation of a photo- z PDF, thus it would be expected to also yield an accurate representation for $N(z)$. Finally, in the last panel we have the results for the sparse basis representation where the number of bases used is defined to be the same as required for the multi-Gaussian fitting method. As seen previously with the distribution of residuals, we see that, with this direct cosmological application, we recover the original $N(z)$, by using the same number of values to represent the photo- z PDF, more accurately than with other techniques.

We present the fractional percentile error between the original $N(z)$ and the reconstructed $N(z)$ computed by using Equation 17 for different photo- z PDF representation techniques in Figure 10. As also seen in Figure 6, we see that as the number of bases increases for the sparse representation (shown in blue dots) the accuracy of the reconstruction also improves, but here we focus on the error in the reconstruction of $N(z)$. Since additional bases will produce a more accurate photo- z PDF representation, we also expect a more accurate $N(z)$ reconstruction when the number of bases increases. For comparison, we also show the multi-Gaussian fitting (red triangle) and the sparse representation (black star) where the number of bases matches the multi-Gaussian fitting value.

We observe that, by using only fifteen values, we can reconstruct the $N(z)$ distribution to an accuracy of 99.9% as

Table 1. The fractional percentile error between the original $N(z)$ and a reconstructed $N(z)$ computed by using the sparse basis representation, single and multi-Gaussian fitting, and all of the single point photo- z techniques described in Section 3.1. We also list the number of values required to represent the photo- z PDF. The table is sorted in ascending order by the percentile error.

Method	Values per PDF	Error [%]
sparse rep. fixed	40	0.05545
sparse rep. fixed	35	0.05551
sparse rep. fixed	30	0.05611
sparse rep. fixed	25	0.05750
sparse rep. fixed	20	0.06829
sparse rep. fixed	15	0.08729
sparse rep. same MG	14	0.12930
sparse rep. fixed	10	0.13440
multi-Gaussian	14	0.19779
sparse rep. fixed	5	0.31113
Monte Carlo	1	0.37294
single Gaussian	2	2.19095
Median PDF	1	6.63550
Mean PDF	1	7.47077
Mode PDF	1	13.24271

measured with respect to the original distribution. In addition, this result changes only slightly when we limit our representation to ten bases. We also see that the error values are slightly better than we saw when reconstructing the individual photo- z PDFs, because computing the $N(z)$ smooths over the individual photo- z PDFs, thereby reducing the impact from small discrepancies in individual photo- z PDFs that might result from using a specific functional basis. If we increase our representation to use forty bases, we can reconstruct the $N(z)$ distribution to nearly 99.95%, but the decrease in the error, however, does not change significantly once we have used approximately twenty-five bases, suggesting there are diminishing returns.

In Section 3.1, we introduced several different individual photo- z estimates that are widely used, including the mean, the mode, and the median of the photo- z PDF, and Monte Carlo sampling from the cumulative photo- z PDF. These single estimates show an even larger fractional error than visible on the vertical axis shown in Figure 10, and are thus presented in Table 1, which summarize the results from all of the methods presented herein, including the number of values required by the representation method and the fractional percentile error for that method in reconstructing the original $N(z)$.

The entries in Table 1 are presented in ascending order by the size of this fractional percentile error. From these entries, we see that the single Gaussian model has, on average, a reconstruction error of 2.2% while the single value estimates all have reconstruction errors over 6%. The Monte Carlo sampling method provides the best reconstruction results when using a single photo- z estimate with an error of about 0.4%, which is comparable to a sparse representation that uses five bases. As mentioned previously, however, this technique does not provide accurate individual photo- z estimates. We also observe that the difference when using thirty, thirty-five, or even forty bases is very small, although it is bigger than the resolution in the discretization scheme; thus our proposed discretization method does not impact these results and we can safely represent each basis function by using a single four-byte integer.

The integration over the dictionary of bases, as shown in

Equation 15, can also be used to compute $N(z)$ over different redshift bins. In this case, the integration can be performed by using the bases and subsequently applying Equation 17 when using the sparse basis representation. Furthermore, we can extend this approach to analyze multiple photo- z PDFs for each galaxy, where they are each represented by the same dictionary. This would prove useful when a survey has stored photo- z PDFs for the same galaxy by using different galaxy spectral templates. Thus, a scientist could either compute an $N(z)$ by using a single, per-galaxy *best* template or compare different $N(z)$ that are computed by using different template combinations. Since the integrals could all be precomputed, the only new computation is for the basis coefficients for each galaxy, dramatically reducing the overall computational demands.

For example, we might have different (or even updated) priors for different galaxy types in a survey. We can quickly apply these new priors to the precomputed dictionary integrals and recover the results for each galaxy given their basis coefficients in an efficient manner. Alternatively, one might want to minimize over the galaxy type under certain restrictions, which can be applied over the precomputed integrals of the dictionary of bases. The minimization problem subsequently becomes a simple task of selecting the minimum or maximum sum over the coefficients. As should be evident from these examples, there exist a number of different applications where our proposed sparse basis representation not only reduces the overall storage requirements, often significantly, but also reduces the computational requirements for cosmological analyses.

6 CONCLUSIONS

In this paper, we have presented different techniques to represent and efficiently store photo- z PDFs, which have been shown to convey significantly more information than a single photo- z estimate. As we enter the era of precision cosmology, the growth of large, dense photometric surveys has created an unmet need to quantify and manage these probabilistic values for hundreds of millions to billions of galaxies. Specifically, we have introduced the use of a sparse basis representation that uses a dictionary of Gaussian functions and Voigt profiles, which have extended wings, to accurately and efficiently represent each photo- z PDF. We minimize the number of required bases while maintaining a high accuracy by using an Orthogonal Matching Pursuit algorithm, which provides a unique set of bases for each photo- z PDF while minimizing the residual between the original and final photo- z PDF. This algorithm is publicly available⁶

We use photometric data from the CFHTLenS survey to compute photo- z PDFs by using our TPZ code, producing PDFs with two hundred points and a redshift resolution of $\delta z = 0.011$. By using these PDFs, we demonstrate that our proposed sparse basis representation reconstructs a more accurate PDF than other techniques, include a multi-Gaussian fitting approach with a flexible number of parameters based on the number of peaks in each PDF. If we use the exact same number of parameters with our sparse representation as used by the multi-Gaussian fitting, we found that the sparse basis

representation results are superior with the additional benefit that each basis or parameter can be stored using a single integer. We also showed that, with a fixed number of bases, we could achieve both a highly accurate PDF that also has a large compression ratio. As a specific example, we found that by using only ten (twenty) values per photo- z PDF, we could reconstruct a photo- z PDF at over a 99.5% accuracy with a compression ratio of twenty (ten), providing a significant storage reduction without a loss of information.

We quantified the number of bases required within the sparse representation dictionary, specifically finding that $(\Delta z / \delta z)^2$ bases are sufficient to represent the galaxy photo- z PDFs in our CFHTLenS test sample, where Δz is the overall redshift range and δz is the photometric redshift PDF resolution. If the number of points in the original PDF is approximately 200–250, we can use a dictionary with fewer than 2^{16} bases, which results in an accurate PDF reconstruction while only requiring a single sixteen-bit integer to store the basis index in the dictionary. Furthermore, since the bases themselves are normalized, all basis coefficients are less than unity by definition; and, since the photo- z PDFs are also normalized, we only need to retain the relative amplitudes of each basis function.

Therefore, we can independently rescale the coefficients for each galaxy to their maximum value and subsequently represent them by using a discretized range containing 2^{15} values. This will provide a resolution less than 10^{-5} , and since we set the maximum value from the most significant basis function, it is always correctly represented. As a result, we can also store the coefficients (sign included) in a separate sixteen-bit integer without losing information. Taken together, we can completely encode a single basis function, both dictionary index and coefficient, in a single four-byte integer, simplifying the data management and significantly reducing the data storage and reconstruction computational requirements.

Of course the results we have presented will depend on the quality of the photo- z PDFs to which they are applied, which themselves depend on the details of the photo- z algorithm that generated them. As would naively be expected, single peaked photo- z PDFs are most accurately reconstructed by using either a multi-Gaussian fitting or a sparse basis representation, where only five points per photo- z PDF is sufficient to achieve a 99% accurate reconstruction. Furthermore, we also demonstrated that, as a simple cosmological application of our photo- z PDF reconstruction, we could accurately recover the underlying $N(z)$ distribution. In particular, we recovered the $N(z)$ of our CFHTLenS test sample to an accuracy of 99.87% by using only ten points per photo- z PDF.

Given their compact nature and the fact that they are predetermined, we showed that we could obtain the sample $N(z)$ by integrating the bases over the sample redshift range and later multiplying by the basis coefficients, which can also be prefactored, thereby significantly reducing the number of required integrations. This same principle can be applied to other linear combinations of photo- z PDFs or to more complex analyses if they can be expressed in terms of the underlying bases. This later topic is the subject of future work. Overall, these results are very promising, as current and future photometric surveys will produce up to tens of billions of photo- z PDFs. Our proposed approach will either allow a reduction in the overall storage requirements or increase the number of photo- z PDFs that can be persistently maintained for each galaxy without increasing the required amount of storage. As

⁶ <http://lcdm.astro.illinois.edu/code/pdfz.html>

a result, this new approach will enable science that would have otherwise been difficult or impossible to accomplish.

ACKNOWLEDGEMENTS

The authors thank the referee for a prompt and careful reading of the manuscript and for comments that improved this work. RJB and MCK acknowledge support from the National Science Foundation Grant No. AST-1313415. MCK has been supported by the Computational Science and Engineering (CSE) fellowship at the University of Illinois at Urbana-Champaign. RJB has been supported in part by the Institute for Advanced Computing Applications and Technologies faculty fellowship at the University of Illinois.

This work also used resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

This work is based on observations obtained with MegaPrime/MegaCam, a joint project of CFHT and CEA/IRFU, at the Canada-France-Hawaii Telescope (CFHT) which is operated by the National Research Council (NRC) of Canada, the Institut National des Sciences de l'Univers of the Centre National de la Recherche Scientifique (CNRS) of France, and the University of Hawaii. This research used the facilities of the Canadian Astronomy Data Centre operated by the National Research Council of Canada with the support of the Canadian Space Agency. CFHTLenS data processing was made possible thanks to significant computing support from the NSERC Research Tools and Instruments grant program.

REFERENCES

- Abdalla F. B., Banerji M., Lahav O., Rashkov V., 2011, *MNRAS*, 417, 1891
- Abrahamse A., Knox L., Schmidt S., Thorman P., Tyson J. A., Zhan H., 2011, *ApJ*, 734, 36
- Abramowitz M., Stegun I., 1972, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, Applied mathematics series. Dover Publications
- Ahn C. P. et al., 2013, *ArXiv e-prints* : 1307.7735
- Benítez N., 2000, *ApJ*, 536, 571
- Bordoloi R., Lilly S. J., Amara A., 2010, *MNRAS*, 406, 881
- Bovy J. et al., 2011, *ApJ*, 729, 141
- Bovy J. et al., 2012, *ApJ*, 749, 41
- Breiman L., 2001, *Machine Learning*, 45, 5
- Carnero A., Sánchez E., Crocce M., Cabré A., Gaztañaga E., 2012, *MNRAS*, 419, 1689
- Carrasco Kind M., Brunner R. J., 2013, *MNRAS*, 432, 1483
- Carrasco Kind M., Brunner R. J., 2014, *MNRAS*, 438, 3409
- Carrasco Kind M., Brunner R. J., 2014b, *ArXiv e-prints* : 1403.0044
- Collister A. A., Lahav O., 2004, *PASP*, 116, 345
- Davis M. et al., 2007, *ApJL*, 660, L1
- Dawson K. S. et al., 2013, *AJ*, 145, 10
- Erben T. et al., 2013, *MNRAS*, 433, 2545
- Garilli B. et al., 2014, *A&A*, 562, A23
- Garilli B. et al., 2008, *A&A*, 486, 683
- Gwyn S. D. J., 2012, *AJ*, 143, 38
- Heymans C. et al., 2012, *MNRAS*, 427, 146
- Hildebrandt H. et al., 2010, *A&A*, 523, A31
- Hildebrandt H. et al., 2012, *MNRAS*, 421, 2355

- Högbom J. A., 1974, *A&AS*, 15, 417
- Ilbert O. et al., 2006, *A&A*, 457, 841
- Jee M. J., Tyson J. A., Schneider M. D., Wittman D., Schmidt S., Hilbert S., 2013, *ApJ*, 765, 74
- Le Fèvre O. et al., 2005, *A&A*, 439, 877
- Lemire D., Boytsov L., 2012, *CoRR*, abs/1209.2137
- Mallat S. G., Zhang Z., 1993, *IEEE Transactions on Signal Processing*, 41, 3397
- Mandelbaum R. et al., 2008, *MNRAS*, 386, 781
- Myers A. D., White M., Ball N. M., 2009, *MNRAS*, 399, 2279
- Newman J. A. et al., 2013a, *ApJS*, 208, 5
- Sánchez C. et al., 2014, in preparation
- Sheldon E. S., Cunha C. E., Mandelbaum R., Brinkmann J., Weaver B. A., 2012, *ApJS*, 201, 32
- Wittman D., 2009, *ApJL*, 700, L174
- York D. G. et al., 2000, *AJ*, 120, 1579

This paper has been typeset from a $\mathrm{T}_{\mathrm{E}}\mathrm{X}/\mathrm{L}^{\mathrm{A}}\mathrm{T}_{\mathrm{E}}\mathrm{X}$ file prepared by the author.